Research Report

Jackknifing Techniques for Evaluation of Equating Accuracy

Shelby J. Haberman

Yi-Hsuan Lee

Jiahe Qian

December 2009

ETS RR-09-39

Listening. Learning. Leading.®

Jackknifing Techniques for Evaluation of Equating Accuracy

Shelby J. Haberman, Yi-Hsuan Lee, and Jiahe Qian
ETS, Princeton, New Jersey

December 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Grouped jackknifing may be used to evaluate the stability of equating procedures with respect to sampling error and with respect to changes in anchor selection. Properties of grouped jackknifing are reviewed for simple-random and stratified sampling, and its use is described for comparisons of anchor sets. Application is made to examples of item response theory (IRT) true-score equating in which two-parameter logistic and general partial credit models are employed.

Key words: True-score equating, generalized partial credit model, two-parameter logistic model

Acknowledgments

Frederic Robin and Jill Carey greatly contributed to this report by providing software and data access. Any opinions expressed in this report are those of the authors and not necessarily those of ETS.

Table of Contents

Page

1 The Traditional and the Grouped Jackknife....2

1.1 Weights....4

1.2 Delete-1 Jackknifing....6

1.3 Grouped Jackknifing....8

1.4 Grouped Jackknifing for Stratified Random Samples....12

1.5 Jackknifing Comparisons....16

1.6 Randomly Selected Estimates....21

1.7 Overlapping Anchor Sets....22

2 IRT True-Score Equating....28

3 Example....31

4 Conclusions....32

References....36

Equating of test forms involves sampling of examinees, so that random equating errors are introduced through estimation of equating parameters. When anchor items are employed in equating and when classical equating assumptions apply, the choice of anchor items should have minimal effect on the equating process. In the real world, it is not necessarily true that choice of anchor items has minimal effect. To evaluate variability in equating due to sampling error and variability of equating due to selection of anchor items, jackknifing may be employed. This report illustrates use of jackknifing in the case of IRT true score equating, but jackknifing may be employed with other approaches as well.

Jackknifing is a commonly employed statistical technique for estimation of variances of sample statistics (Quenouille, 1956; Tukey, 1958; Miller, 1964). It may be employed to obtain approximate confidence intervals for population measures of interest. Applications of jackknifing commonly involve cases in which it is difficult to apply the δ-method (Rao, 1973, p. 388) to estimate variances. Given the large number of steps involved in IRT true-score equating, the δ-method is challenging to apply; however, the grouped jackknifing approach (Miller, 1964) is readily used to study sampling errors associated with conversions of test scores. *Grouped jackknifing* is an example of a resampling method because it employs estimates based on selected subsamples of the observed data. It requires much less computational labor than other resampling methods such as bootstrapping methods (Efron, 1979, 1982), traditional jackknifing (Quenouille, 1956; Tukey, 1958), or delete-d versions of the jackknife in which $d > 1$ (Shao & Wu, 1989).

Jackknifing may also be employed to examine the stability of IRT true-score equating with respect to the choice of anchor items. This stability can be examined in two distinct fashions. In one case, the effect of a specified change in the anchor set can be studied by examination of the estimated means and standard deviations of the differences between the resulting conversions. In another case, anchor items can be regarded as a sample from a collection of possible anchor items. One then examines both the variability of conversions due to sampling of examinees and the variability of conversions due to selection of anchor items. This latter possibility has been considered previously (Cohen, Johnson, & Angeles, 2001); however, this application of jackknifing requires additional study to justify its use. In addition, within the context of equating, consideration must also be given to the nature of sampling in the case of items. In typical cases, testing programs do not randomly select items, so that inferences may be problematic beyond the anchor items present in the forms under study. This issue will be discussed further in section 4.

Section 1 provides necessary background concerning the grouped jackknife. Section 2 provides background concerning IRT true-score equating. In section 3, jackknifing is applied to assess variability of conversions in two cases in which two forms of a test are linked by IRT true-score equating. Section 4 provides some general observations concerning application of jackknifing to the study of equating.

1 The Traditional and the Grouped Jackknife

The grouped jackknife is an old example of a resampling method (Efron, 1979, 1982). It is primarily of interest when computational cost is a major issue. To explain grouped jackknifing, it is helpful to begin with elementary methods to estimate standard errors and obtain confidence intervals for the population mean and population standard deviation. These examples lead to some simple illustrations of traditional delete-1 jackknifing procedures in which a series of estimates are computed by removing one observation from the sample. The analysis of traditional delete-1 jackknifing then leads to grouped jackknifing in which the observations are divided into groups and estimates are computed by leaving out one group from the sample.

In discussion of delete-1 jackknifing, the sample mean of independent and identically distributed random variables has a fundamental role. One basic justification of delete-1 jackknifing is the fact that it results in customary inferences concerning the population mean when the sample mean is employed. General justification of delete-1 jackknifing involves a demonstration that the parameter estimates under study are well approximated by sample means. Such approximations are typically available when parameter estimates are differentiable functions of sample means.

To begin, consider the sample mean of the real observations X_i, $1 \leq i \leq n$, $n > 1$, obtained by random sampling with replacement. For example, the X_i might be raw scores of examinees for a particular test administration, where the examinees are regarded as a sample from a hypothetical infinite population of potential examinees. Let the X_i be random variables with common mean μ and common variance $\tau^2 > 0$. The assumption of random sampling with replacement implies that the X_i are independent and identically distributed. Consider the elementary problem of estimation of the expectation μ by the sample mean

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i.$$

As is well known, $\bar{X}$ has expectation μ and variance $\sigma^2(\bar{X}) = \tau^2/n$. Thus $\bar{X}$ is an unbiased

estimate of μ. In addition, the variance $\sigma^2(\bar{X})$ has a simple unbiased estimate, for the sample variance

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

has expectation τ^2, and $\hat{\sigma}^2(\bar{X}) = s^2/n$, the estimated variance of the sample mean, is an unbiased estimate of $\sigma^2(\bar{X}) = \tau^2/n$, the variance of the sample mean. In addition, the ratio $\hat{\sigma}^2(\bar{X})/\sigma^2(\bar{X})$ converges to 1 with probability 1 as the sample size n becomes large (Shao, 2003, p. 133). The estimated standard error $\hat{\sigma}(\bar{X})$ of the sample mean is the square root of the estimated variance $\hat{\sigma}^2(\bar{X})$ of the sample mean. When the sample size n is large, $(\bar{X} - \mu)/\hat{\sigma}(\bar{X})$ has an approximate standard normal distribution, so that approximate confidence intervals for μ are readily constructed (Scheffé, 1959, p. 355). For any real α such that $0 < \alpha < 1$ and any positive integer ν, let $t_{\nu,\alpha}$ be defined so that α is the probability that a random variable with a t distribution on ν degrees of freedom has absolute value at least as large as $t_{\nu,\alpha}$. In addition, let z_α be defined so that α is the probability that a random variable with a standard normal distribution has absolute value at least as large as $t_{\nu,\alpha}$. Then the customary approximate two-sided confidence interval for μ of level $1 - \alpha$ has lower bound

$$\mu_{L\alpha} = \bar{X} - t_{n-1,\alpha}\hat{\sigma}(\bar{X})$$

and upper bound

$$\mu_{U\alpha} = \bar{X} + t_{n-1,\alpha}\hat{\sigma}(\bar{X}).$$

As the sample size n increases, the probability approaches $1 - \alpha$ that $\mu_{L\alpha} \le \mu \le \mu_{U\alpha}$. In the special case in which the X_i have a common normal distribution, $(\bar{X} - \mu)/\hat{\sigma}(\bar{X})$ has a t distribution on $n - 1$ degrees of freedom, and $(n-1)s^2/\tau^2$ has a chi-squared distribution on $n - 1$ degrees of freedom, so that $1 - \alpha$ is the exact probability that $\mu_{L\alpha} \le \mu \le \mu_{U\alpha}$.

In the discussion of grouped jackknifing and traditional delete-1 jackknifing, comparison of results is made with those obtained by traditional confidence intervals for the population mean. One aspect of this comparison involves expected widths of confidence intervals. These expected widths are not difficult to study in the case of the approximate confidence intervals for the population mean. For a large sample size n, the multiplier $t_{n-1,\alpha}$ is close to z_α. Even for $n = 120$ and $\alpha = 0.05$, $t_{n-1,\alpha} = 1.9801$ and $z_\alpha = 1.9600$. In general, $t_{\nu,\alpha}$ is quite well approximated by $z_\alpha + (z_\alpha + z_\alpha^3)/(4\nu)$ as ν increases (Abramowitz & Stegun, 1965, p. 949). For example, for $\nu = 119$

and $\alpha = 0.05$,

$$z_\alpha + (z_\alpha + z_\alpha^3)/(4\nu) = 1.9799$$

is quite close to $t_{n-1,\alpha} = 1.9801$. In the case of the X_i normally distributed, the expected width $E(\mu_{U\alpha} - \mu_{L\alpha})$ of the confidence interval of level $1 - \alpha$ is readily found. Because $E(s) = \Gamma(n/2)[2/(n-1)]^{1/2}\tau/\Gamma((n-1)/2))$ (Cramér, 1946, p. 383), where Γ denotes the gamma function,

$$E(\mu_{U\alpha} - \mu_{L\alpha}) = 2t_{n-1,\alpha}\Gamma(n/2)[2/n(n-1)]^{1/2}\tau/\Gamma((n-1)/2)).$$

This width is quite close to $z_\alpha\tau/n^{1/2}$ even for n of moderate size. For example, if $n = 120$ and $\alpha = 0.05$, the width of $3.9519\tau/n^{1/2}$ is quite close to $2z_\alpha\tau/n^{1/2} = 3.9199\tau/n^{1/2}$.

These familiar results for the sample mean do not apply even for such simple summary statistics as the sample standard deviation s, the square root of s^2. The sample standard deviation is commonly used to estimate the common standard deviation τ of the observations X_i. Nonetheless, the expectation $E(s)$ of s is not τ, and the variance $\sigma^2(s)$ of s does not have an unbiased estimate. The δ method can be used to study statistical properties of s when the variance υ^2 of $Y_i = [(X_i - \mu)^2 - \tau^2]/(2\tau)$ is finite and positive (Cramér, 1946, p. 353). In this case, as the sample size n increases, s is well approximated by $\tau + \bar{Y}$, where $\bar{Y}$ is the sample mean of the Y_i, $1 \le i \le n$. Let $R = s - \tau - \bar{Y}$ denote the approximation error. As the sample size n increases, the mean squared error $E(R^2)$ is sufficiently small that $E(R^2)/\sigma^2(\bar{Y})$ approaches 0. The mean $E(s)$ approaches τ sufficiently rapidly that $[E(s) - \tau]/\sigma^2(\bar{Y})$ converges to $-\upsilon^2/(2\tau)$. The variance $\sigma^2(s)$ is well approximated by $\sigma^2(\bar{Y})$ in the sense that $[\sigma^2(\bar{Y}) - \sigma^2(s)]/\sigma^2(s)$ approaches 0 as the sample size increases. The Y_i are not observed, but one may approximate Y_i by $\hat{Y}_i = [(X_i - \bar{X})^2 - s^2]/(2s)$ and obtain an estimate $\hat{\sigma}^2(s)$ for $\sigma^2(s)$ equal to the estimated variance of the sample mean for observations $\hat{Y}_i$, $1 \le i \le n$. An approximate confidence interval for τ is based on the observation that $(s - \tau)/\hat{\sigma}(s)$ has an approximate standard normal distribution if the sample size n is large. In addition, the ratio $\hat{\sigma}^2(s)/\sigma^2(s)$ converges in probability to 1 as the sample size n becomes large; that is, for any positive real number ϵ, as the sample size n increases, the probability that $\hat{\sigma}^2(s)/\sigma^2(s)$ differs from 1 by more than ϵ approaches 0.

1.1 Weights

Resampling methods provide an alternative approach to variance estimation. These methods can be described in terms of sampling weights (Efron, 1982, p. 37). For example, consider the

sample mean $\bar{X}$. For each observation i, let $w_i \geq 0$ be an integer weight assigned to sample member i. The weight w_i will represent the number of times sample member i is to be used in computation of an estimate. Let $\mathbf{w}$ denote the n-dimensional weight vector with coordinate i equal to w_i, and let the sum $n[\mathbf{w}] = \sum_{i=1}^{n} w_i$ of the weights be positive. Then one may consider the weighted mean

$$\bar{X}[\mathbf{w}] = \{n[\mathbf{w}]\}^{-1} \sum_{i=1}^{n} w_i X_i.$$

Thus $\bar{X}$ is $\bar{X}[\mathbf{1}]$, where $\mathbf{1}$ is the vector with all coordinates 1. If $w_1 = 0$ and $w_i = 1$ for $i > 1$, then

$$\bar{X}[\mathbf{w}] = (n-1)^{-1} \sum_{i=2}^{n} X_i$$

is the sample mean for the observations X_2 to X_n. In general, $\bar{X}[\mathbf{w}]$ has expectation μ, just as in the case of the original sample mean $\bar{X}$.

Weights can also be used with the sample variance and sample standard deviation. Let $n[\mathbf{w}] > 1$, let

$$s^2[\mathbf{w}] = \{n[\mathbf{w}] - 1\}^{-1} \sum_{i=1}^{n} w_i (X_i - \bar{X}[\mathbf{w}])^2$$

and let $s[\mathbf{w}]$ be the square root of $s^2[\mathbf{w}]$. If all weights w_i are 0 or 1, then $s^2[\mathbf{w}]$ has expectation τ^2. Note that $s^2[\mathbf{1}]$ is the sample variance s^2 of the X_i, $1 \leq i \leq n$, and $s[\mathbf{1}]$ is the corresponding sample standard deviation s. If $w_1 = 0$ and $w_i = 1$ for $i > 1$, then $s^2[\mathbf{w}]$ is the sample variance deviation for the observations X_2 to X_n, and $s[\mathbf{w}]$ is the corresponding sample standard deviation.

In general, estimates $g[\mathbf{w}]$ will be considered for a real parameter γ, where $g[\mathbf{1}]$ will be denoted by g. For the weight vectors $\mathbf{w}$ under study, the essential requirements are that $g[\mathbf{w}]$ have finite variance and that independent and identically distribution random variables Y_i, $1 \leq i \leq n$, with mean 0 and variance $\upsilon^2 > 0$ exist such that the estimates $g[\mathbf{w}]$ are well approximated by $\gamma + \bar{Y}[\mathbf{w}]$, where the weighted mean

$$\bar{Y}[\mathbf{w}] = \{n[\mathbf{w}]\}^{-1} \sum_{i=1}^{n} w_i Y_i$$

(Shao & Wu, 1989). In the case of $\bar{X}[\mathbf{w}]$, $Y_i = X_i - \mu$. In the case of $s[\mathbf{w}]$, the requirement is met with $Y_i = [(X_i - \mu)^2 - \tau^2]/\tau$. The approximation requirements involve the approximation error

$$R = g - \gamma - \bar{Y} \tag{1}$$

for the complete sample and the approximation error

$$R[\mathbf{w}] = g[\mathbf{w}] - \gamma - \bar{Y}[\mathbf{w}] \tag{2}$$

for the weight vector $\mathbf{w}$ with integer weight $w_i \geq 0$ assigned to sample member i.

1.2 Delete-1 Jackknifing

In traditional delete-1 jackknifing (Quenouille, 1956; Shao & Wu, 1989; Tukey, 1958), weight vectors $\mathbf{w}(j)$, $1 \leq j \leq n$, are employed to compute n sample statistics. These weight vectors correspond to samples in which one member is omitted. Thus the weight vector $\mathbf{w}(j)$ provides a weight $w_i(j) = 1$ to each sample member i not equal to j, but the weight $w_j(j)$ for sample member j is 0. For sample member j, the delete-1 estimate $g[\mathbf{w}(j)]$ corresponds to an estimate of γ based on the observed X_i for all sample members i except j. For example, $\mathbf{w}(1)$ has coordinate $w_1(1)$ equal to 0 and coordinates $w_i(1) = 1$ for $i \geq 2$, so that $g[\mathbf{w}(1)]$ is the estimate based on the observations X_i, $i > 1$. The average delete-1 estimate is then

$$\bar{g} = n^{-1} \sum_{j=1}^{n} g[\mathbf{w}(j)].$$

The jackknife variance estimate for $\sigma^2(g)$ is

$$\hat{\sigma}_J^2(g) = \frac{n-1}{n} \sum_{j=1}^{n} \{g[\mathbf{w}(j)] - \bar{g}\}^2.$$

The delete-1 jackknife has desirable large-sample properties when two conditions both hold. The first condition is that the mean squared approximation error $E(R^2)$ associated with the complete sample is sufficiently small so that

$$E(R^2)/\sigma^2(\bar{Y}) \to 0 \tag{3}$$

as the sample size n becomes large. The second condition requires that the difference $R - R[\mathbf{w}(j)]$ between the approximation errors R for the complete sample and $R[\mathbf{w}(j)]$ for the sample with member j omitted is sufficiently small so that

$$\max_{1 \leq j \leq n} E(\{R - R[\mathbf{w}(j)]\}^2)/[\sigma^2(\bar{Y})]^2 \to 0 \tag{4}$$

as the sample size n increases (Shao & Wu, 1989). Under these conditions, the sample variance $\sigma^2(g)$ is well approximated by $\sigma^2(\bar{Y}) = v^2/n$ in the sense that $\sigma^2(g)/\sigma^2(\bar{Y})$ converges to 1 as the sample size n becomes large. The bias $E(g) - \gamma$ is sufficiently small so that $[E(g) - \gamma]/\sigma(g)$ converges to 0 as the sample size n increases. The approximation $\hat{\sigma}_J^2(g)$ to the sample variance

$\sigma^2(g)$ is sufficiently accurate so that $\hat{\sigma}_J^2(g)/\sigma_J^2(g)$ converges in probability to 1 as the sample size n increases, and the ratio $(g-\gamma)/\hat{\sigma}_J(g)$ has an approximate standard normal distribution.

Approximate confidence intervals for γ are readily constructed. For consistency with practice for the sample mean, for real α such that $0<\alpha<1$, let the lower bound of the approximate confidence interval for γ of level $1-\alpha$ be

$$\gamma_{JL\alpha} = g - t_{n-1,\alpha}\hat{\sigma}_J(g),$$

and let the upper bound be

$$\gamma_{JU\alpha} = g + t_{n-1,\alpha}\hat{\sigma}_J(g).$$

Then the probability that $\gamma_{JL\alpha} \le \gamma \le \gamma_{JU\alpha}$ approaches $1-\alpha$ as the sample size n increases.

In the case of the sample mean, (3) and (4) hold trivially if $Y_i = X_i - \mu$, $\gamma = \mu$, $\upsilon^2 = \tau^2$, and $g = \bar{X}$, for R and $R[\mathbf{w}(j)]$ are 0. Delete-1 jackknifing leads to conventional inferences concerning the population mean. The average of the $\bar{X}[\mathbf{w}(j)]$, $1 \le j \le n$, is the original sample mean $\bar{X}$, and

$$\hat{\sigma}_J^2(\bar{X}) = \frac{n-1}{n}\sum_{j=1}^{n}(\bar{X}[\mathbf{w}(j)] - \bar{X})^2 = \hat{s}^2(\bar{X})$$

(Efron, 1982, pp. 6, 13). Thus jackknifing simply leads to the conventional estimate of the variance of the sample mean. In addition, the jackknife confidence bounds $\mu_{JL\alpha}$ and $\mu_{JU\alpha}$ satisfy $\mu_{JL\alpha} = \mu_{JL\alpha}$ and $\mu_{JU\alpha} = \mu_{U\alpha}$.

In the case of the sample standard deviation, (3) and (4) may be shown to hold if $\gamma = \tau$, $g = s$, and $Y_i = [(X_i-\mu)^2 - \tau^2]/(2\tau)$. Jackknifing yields a different estimate of the variance of s than the one obtained previously by use of the δ method. Let $n > 2$. One has

$$\bar{s} = n^{-1}\sum_{j=1}^{n} s[\mathbf{w}(j)],$$

and

$$\hat{\sigma}_J^2(s) = \frac{n-1}{n}\sum_{j=1}^{n}(s[\mathbf{w}(j)] - \bar{s})^2.$$

The ratio $\hat{\sigma}_J^2(s)/\sigma^2(s)$ converges in probability to 1 as the sample size n becomes large, $(s-\tau)/\hat{\sigma}_J(s)$ converges in distribution to a random variable with a standard normal distribution, and, for real α such that $0<\alpha<1$, the approximate confidence interval for τ of level $1-\alpha$ has lower bound

$$\tau_{JL\alpha} = s - t_{n-1,\alpha}\hat{\sigma}_J(s)$$

and upper bound

$$\tau_{JU\alpha} = s + t_{n-1,\alpha}\hat{\sigma}_J(s).$$

Computations are somewhat easier than may at first appear to be the case, for

$$s^2[\mathbf{w}(j)] = (n-2)^{-1}[(n-1)s^2 - n(X_j - \bar{X})^2/(n-1)]$$

(Draper & Smith, 1998, p. 208). As the sample size n increases, the probability approaches $1-\alpha$ that $\tau_{JL\alpha} \le \tau \le \tau_{JU\alpha}$.

It is possible to demonstrate that the requirements for delete-1 jackknifing are met for the equating applications under study under some possible sampling models; however, computation of this jackknife estimate of the variance is impractical in the equating examples considered. Thousands of observations are involved, and the computer programs used in calculations do not permit any simplification of calculations comparable to that achievable for the sample standard deviation. As a consequence, other resampling approaches must be considered.

1.3 Grouped Jackknifing

The grouped jackknife (Miller, 1964) is a less computationally intensive resampling alternative to the traditional jackknife. In this approach, the n observations are divided into $k \le n$ disjoint groups G_j, $1 \le j \le k$, with approximately equal numbers of members. In the simplest case, the sample size n is a multiple of k, so that each group G_j can be selected to have $n(G_j) = n/k$ members. For example, if $n = 100$ and $k = 10$, then one might have G_1 contain observations 1 to 10, G_2 contain observations 11 to 20, and G_{10} contain observation 91 to 100. More generally, the groups can always be chosen so that $|n(G_j) - n/k|$ is less than 1. For example, if n is 101 and k is 10, then G_1 to G_9 can be defined as in the case of $k = 10$ and $n = 100$; however, G_{10} may now be defined so that G_{10} contains observations 91 to 101. The weight vectors $\mathbf{w}_G(j)$, $1 \le j \le k$, are selected so that $\mathbf{w}_G(j)$ has ith coordinate $w_{iG}(j)$ equal to 1 if observation i is not in group G_j. Coordinate $w_{iG}(j)$ is 0 if i is in group G_j. For example, the delete-$n(G_j)$ sample mean $\bar{X}[\mathbf{w}_G(j)]$ is the sample mean of the X_i for observations i not in group G_j. With grouped jackknifing, the variance estimate

$$\hat{\sigma}^2_G(g) = \frac{k-1}{k}\sum_{j=1}^{k}(g[\mathbf{w}_G(j)] - \bar{g}_G)^2,$$

where the average of the delete-$n(G_j)$ estimates $g[\mathbf{w}(j)]$, $1 \le j \le k$, is

$$\bar{g}_G = k^{-1} \sum_{j=1}^{k} g[\mathbf{w}_G(j)].$$

For $0 < \alpha < 1$, the approximate confidence interval of level $1 - \alpha$ for γ has lower bound

$$\gamma_{GL\alpha} = g - t_{k-1,\alpha} \hat{\sigma}_G(g)$$

and upper bound

$$\gamma_{GU\alpha} = g + t_{k-1,\alpha} \hat{\sigma}_G(g).$$

Traditional delete-1 jackknifing is obtained in the special case of $k = n$, for each $G(j)$ contains only one member of the sample, and $\mathbf{w}_G(j) = \mathbf{w}(j)$. When $k < n$, grouped jackknifing is different from delete-1 jackknifing even in simple cases such as estimation of the variance of the sample mean. In the applications under study, the number k is fixed by restrictions on computational resources. For example, in the equating problems under study, k is 120 no matter how large the sample size n may be. This restriction greatly reduces computational labor relative to alternatives. Delete-1 jackknifing requires n subsamples. Delete-d jackknifing, $1 < d < n - 1$, requires all subsamples in which d members are omitted from the original sample (Shao & Wu, 1989), so that even more subsamples are required than for delete-1 jackknifing. For the applications under study, no obvious gain is achieved from use of $k = 120$ bootstrap samples rather than the grouped jackknife.

The behavior of grouped jackknifing is easiest to examine if n/k is an integer and if g is the sample mean $\bar{X}$. In this case, results can be regarded as quite satisfactory, although there is some loss in terms of width of confidence intervals if $k < n$. Nonetheless, this loss is small for k of moderate size. To verify this claim, let $\mathbf{v}_G(j) = \mathbf{1} - \mathbf{w}_G(j)$, so that $\mathbf{v}_G(j)$ has coordinate $v_{iG} = 1$ for sample member i is in G_j and $v_{iG} = 0$ if sample member i is not in G_j. Thus $\bar{X}[\mathbf{w}_G(j)]$ is the average of the X_i for sample members i not in group G_j, and $\bar{X}[\mathbf{v}_G(j)]$ is the average of the X_i for sample members i in group G_j. Because each group G_j has n/k members, $n[\mathbf{w}_G(j)] = n(k-1)/k$, $n[\mathbf{v}_G(j)] = n/k$, and

$$(k-1)\bar{X}[\mathbf{w}_G(j)] + \bar{X}[\mathbf{v}_G(j)] = \bar{X}.$$

The sample mean $\bar{X}$ is both the average of the delete-n/k sample means $\bar{X}[\mathbf{w}_G(j)]$, $1 \le j \le k$, and the average of the sample means $\bar{X}[\mathbf{v}_G(j)]$ for sample members in group G_j, $1 \le j \le k$. As in

the traditional jackknife, it is readily checked that

$$\hat{\sigma}_G^2(\bar{X}) = [k(k-1)]^{-1} \sum_{j=1}^{k} \{\bar{X}[\mathbf{v}_G(j)] - \bar{X}\}^2.$$

Because the G_j are disjoint groups and the X_i are independent and identically distributed, the sample means $\bar{X}[\mathbf{v}_G(j)]$, $1 \le j \le k$, are independent and identically distributed with common mean μ and common variance $\tau^2/(n/k)$. Thus $\hat{\sigma}_G^2(\bar{X})$ has mean $\sigma_G^2(\bar{X})$, so that $\hat{\sigma}_G^2(\bar{X})$ is an unbiased estimate of the variance of $\bar{X}$. For $0 < \alpha < 1$, the approximate confidence interval for μ of level $1-\alpha$ has lower bound

$$\mu_{GL\alpha} = \bar{X} - t_{k-1,\alpha}\hat{\sigma}_G(\bar{X})$$

and upper bound

$$\mu_{GU\alpha} = \bar{X} + t_{k-1,\alpha}\hat{\sigma}_G(\bar{X}).$$

If the X_i are normally distributed, then the $\bar{X}[\mathbf{v}_G(j)]$ are also normally distributed, so that $(k-1)\hat{\sigma}_G^2(\bar{X})/\sigma_G^2(\bar{X})$ has a chi-square distribution on $k-1$ degrees of freedom and $(\bar{X}-\mu)/\hat{\sigma}_G(\bar{X})$ has a t distribution on $k-1$ degrees of freedom. This exact result is not available if bootstrapping is used. For $0 < \alpha < 1$, $1-\alpha$ is the probability that $\mu_{GL\alpha} \le \mu \le \mu_{GU\alpha}$. Results are quite different from the traditional jackknife to the extent that $\hat{\sigma}_G^2(\bar{X})/\sigma^2(\bar{X})$ does not converge in probability to 1 as the sample size becomes large (Shao & Wu, 1989). Because a chi-square random variable with $k-1$ degrees of freedom has mean $k-1$ and variance $2(k-1)$, the ratio $\hat{\sigma}_G^2(\bar{X})/\sigma^2(\bar{X})$ has mean 1, variance $2/(k-1)$, and standard deviation $[2/(k-1)]^{1/2}$ for all sample sizes. In the case of $k = 120$ considered in this report, the standard deviation $[2/119]^{1/2} = 0.13$ is certainly not negligible, so that variability of $\hat{\sigma}_G^2(\bar{X})$ cannot be ignored. Despite the variability of $\hat{\sigma}_G^2(\bar{X})$, the impact on confidence intervals for μ is relatively small if $\hat{\sigma}_G(\bar{X})$ is used instead of $\hat{\sigma}(\bar{X})$. Recall that in traditional jackknifing, the expected width of the confidence interval at level $1-\alpha$ is

$$E(\mu_{JU\alpha} - \mu_{JL\alpha}) = E(\mu_{U\alpha} - \mu_{L\alpha}) = 2t_{n-1,\alpha}\Gamma(n/2)[2/n(n-1)]^{1/2}\tau/\Gamma((n-1)/2)).$$

A very similar argument may be employed to show that the expected width of the confidence interval at level $1-\alpha$ from grouped jackknifing is

$$E(\mu_{GU\alpha} - \mu_{GL\alpha}) = 2t_{k-1,\alpha}\Gamma(k/2)[2/n(k-1)]^{1/2}\tau/\Gamma((k-1)/2)).$$

As the sample size becomes large, the ratio

$$\frac{E(\mu_{GU\alpha} - \mu_{GL\alpha})}{E(\mu_{JU\alpha} - \mu_{JL\alpha})} = \frac{t_{k-1,\alpha}\Gamma(k/2)n^{1/2}\Gamma((n-1)/2)}{t_{n-1,\alpha/2}\Gamma(n/2)(k-1)^{1/2}\Gamma(n/2)}$$

approaches

$$\frac{2^{1/2}t(k-1,\alpha)\Gamma(k/2)}{(k-1)^{1/2}\Gamma((k-1)/2)}.$$

For $k = 120$, this ratio is 1.0082, a value only slightly greater than 1.

To obtain more general results concerning grouped jackknifing for the sample mean, apply the central limit theorem and the Mann-Wald theorem (Rao, 1973, p. 124). It follows that, as the sample size n becomes large, the distribution of $(k-1)\hat{\sigma}^2_G(\bar{X})/\sigma^2_G(\bar{X})$ has an approximate chi-square distribution on $k-1$ degrees of freedom, and $(\bar{X}-\mu)/\hat{\sigma}_G(\bar{X})$ has an approximate t distribution on $k-1$ degrees of freedom. Thus, even for large samples, it remains the case that $\hat{\sigma}^2_G(\bar{X})$ has limited accuracy as an estimate of $\sigma^2(\bar{X})$. Nonetheless, as the sample size n becomes large, the probability approaches $1-\alpha$ that $\mu_{GL\alpha} \le \mu \le \mu_{GU\alpha}$. As long as $\tau + (X_i-\mu)^2/(2\tau)$ has finite variance, it remains the case that

$$\frac{E(\mu_{GU\alpha}-\mu_{GL\alpha})}{E(\mu_{JU\alpha}-\mu_{JL\alpha})}$$

approaches

$$\frac{2^{1/2}t(k-1,\alpha)\Gamma(k/2)}{(k-1)]^{1/2}\Gamma((k-1)/2)}$$

as the sample size increases.

The basic results for the sample mean extend readily to more general estimates g. Define the Y_i as in the case of the traditional jackknife so that $g[\mathbf{w}]$ is approximated by $\gamma + \bar{Y}$, the Y_i are independent and identically distributed, and the Y_i have common mean 0 and common variance $v^2 > 0$. Conditions for large-sample approximations are a bit weaker than in the traditional jackknife (Shao & Wu, 1989). It suffices to have (3) hold and to have

$$\max_{1\le j\le k} E(\{R - R[\mathbf{w}_G(j)]\}^2)/\sigma^2(\bar{Y}) \to 0 \tag{5}$$

hold as the sample size n increases (Shao & Wu, 1989). Under these conditions, it remains true that the variance $\sigma^2(g)$ is well approximated by the variance $\sigma^2(\bar{Y})$ in the sense that $\sigma^2(g)/\sigma^2(\bar{Y})$ converges to 1 as the sample size n becomes large. In addition, the bias $E(g)-\gamma$ is sufficiently small that $[E(g)-\gamma]/\sigma(g)$ converges to 0 as n becomes large. As in the case of the grouped jackknife of the sample mean, the ratio $\hat{\sigma}^2_G(g)/\sigma^2(g)$ does not converge in probability to 1 as the sample size increases. Instead, the distribution of $(k-1)\hat{\sigma}^2_G(g)/\sigma^2(g)$ has an approximate chi-square distribution on $k-1$ degrees of freedom, and $(g-\gamma)/\hat{\sigma}_G(g)$ has an approximate t

distribution on $k-1$ degrees of freedom. It follows that, as the sample size n increases, the probability approaches $1-\alpha$ that $\gamma_{GL\alpha} \leq \gamma \leq \gamma_{GU\alpha}$.

Grouped jackknifing can be used for some equating designs to evaluate equating error; however, in the cases under study in this report, it is probably appropriate to consider an adaptation of grouped jackknifing to stratified random sampling.

1.4 Grouped Jackknifing for Stratified Random Samples

Jackknifing is often applied when sampling is much more complex than in the case of simple random sampling (Wolter, 1985). A variety of possible approaches exist. In the analysis of equating under study, grouped jackknifing is applied to statistics computed from data from two independent stratified random samples. Similar studies can be performed on data from several independent random samples. To illustrate the approach, consider the case of $H \geq 2$ populations. For each population h, consider $n_h \geq 2$ observations X_{ih}, $1 \leq i \leq n_h$, derived by simple random sampling with replacement. A basic requirement for grouped jackknifing for the stratified case is that it works in a satisfactory manner when a linear combination of sample means is used to estimate a corresponding linear combination of population means. As in the case of grouped jackknifing or delete-1 jackknifing for simple random sampling, further use of jackknifing can then be justified by consideration of parameter estimates well approximated by linear combinations of sample means.

To examine the estimation problem for sample means, let the independent random variables X_{ih} have mean μ_h and variance $\tau_h^2 > 0$ for $1 \leq i \leq n_h$ and $1 \leq h \leq H$. Consider estimation of a linear combination

$$\gamma = \sum_{h=1}^{H} c_h \mu_h \tag{6}$$

of the means μ_h, $1 \leq h \leq H$, for some real numbers c_h, $1 \leq h \leq H$. For example, if $c_h = H^{-1}$ for each population h, then γ is the average $\bar{\mu}$ of the population means μ_h. The conventional estimate of γ is the linear combination

$$g = \sum_{h=1}^{H} c_h \bar{X}_h \tag{7}$$

of the sample means

$$\bar{X}_h = n_h^{-1} \sum_{i=1}^{n_h} X_{ih}.$$

The mean of g is γ, so that g is unbiased, and the variance of g is

$$\sigma^2(g) = \sum_{h=1}^{H} c_h^2 \tau_h^2 / n_h.$$

One may estimate the variance $\sigma^2(g)$ of γ by

$$\hat{\sigma}^2(g) = \sum_{h=1}^{H} c_h^2 s_h^2 / n_h,$$

where the sample variance of the X_{ih}, $1 \le i \le n_h$, is

$$s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (X_{ih} - \bar{X}_h)^2.$$

The ratio $\hat{\sigma}^2(g)/\sigma^2(g)$ converges in probability to 1 if each sample size n_h becomes large, and $(g - \gamma)/\hat{\sigma}(g)$ has an approximate standard normal distribution if all n_h are large.

As in simple random samples, weights can be applied to stratified random samples. Let $\mathbf{w}$ have nonnegative integer coordinates w_{ih}, $1 \le i \le n_h$, $1 \le h \le H$, let $n_h[\mathbf{w}] = \sum_{i=1}^{n_h} w_{ih} > 0$ be the weight sum for sample h, and let

$$\bar{X}_h[\mathbf{w}] = \{n_h[\mathbf{w}]\}^{-1} \sum_{i=1}^{n_h} w_{ih} X_{ih}$$

be the weighted sample mean for sample h. Corresponding to the linear combination g of (7) is the linear combination

$$g[\mathbf{w}] = \sum_{h=1}^{H} c_h \bar{X}_h[\mathbf{w}]. \tag{8}$$

Then $g[\mathbf{w}]$ has mean γ. Similarly, if $n_h[\mathbf{w}] > 1$, then one may let

$$s_h^2[\mathbf{w}] = \{n_h[\mathbf{w}]\} - 1)^{-1} \sum_{i=1}^{n_h} w_{ih} \{X_{ih} - \bar{X}_h[\mathbf{w}]\}^2,$$

so that if each w_{ih} is 0 or 1, then $s_h^2[\mathbf{w}]$ is the sample variance of the observations X_{ih} for which $w_{ih} = 1$. If the w_{ih} are all 0 or 1, then $s_h^2[\mathbf{w}]$ has expectation τ_h^2. If each w_{ih} is 1, then $\bar{X}_h[\mathbf{w}] = \bar{X}_h$, $s_h^2[\mathbf{w}] = s_h^2$, and $g[\mathbf{w}]$ is g.

In the version of grouped jackknifing considered here, for a given positive integer k no greater than the minimum of the sample sizes n_h, $1 \le h \le H$, the sample members drawn from population h are divided into k groups G_{jh}, $1 \le j \le k$, of approximately equal size. If n_h/k is an integer, then each group G_{jh} contains $n(G_{jh}) = n_h/k$ observations. In general, $|n(G_{jh}) - n_h/k|$ is less

than 1. For example, consider $H = 2$, let $k = 10$ groups be taken from for $n_1 = 100$ members of the first sample and $n_2 = 200$ members of the second sample. In this case, G_{11} might be observations 1 to 10 from the first sample, and G_{12} might be observations 1 to 20 from the second sample. One might have G_{91} equal to observations 81 to 90 in the first sample and G_{92} equal to observations 161 to 180 from the second sample. In grouped jackknifing, weight functions $\mathbf{w}_{GS}(j)$, $1 \le j \le k$, are considered such that $\mathbf{w}_{GS}(j)$ has coordinates $w_{ihGS}(j)$, $1 \le i \le n_h$, $1 \le h \le 2$, such that $w_{ihGS}(j)$ is 1 if i is not in G_{jh} and 0 if i is in G_{jh}. For example, $X_1[\mathbf{w}_{GS}(1)]$ is the average of the observations X_{i1} for sample members i from the first sample that are not in group G_{11}. In applications, the standard estimate of a real parameter γ is $g = g[\mathbf{1}_{GS}])$, where $\mathbf{1}_{GS}$ has all coordinates $1_{ihGS} = 1$. In addition, the estimates $g[\mathbf{w}_{GS}(j)]$ are used for variance estimation. It is assumed that g and $g[\mathbf{w}_{GS}(j)]$ have finite variances. At this point, calculations are essentially the same as for the grouped jackknife for simple random sampling with replacement. The variance $\sigma^2(g)$ is estimated by

$$\hat{\sigma}^2_{GS}(g) = \frac{k-1}{k} \sum_{j=1}^{k} (g[\mathbf{w}_{GS}(j)] - \bar{g}_{GS})^2,$$

where

$$\bar{g}_{GS} = k^{-1} \sum_{j=1}^{k} g[\mathbf{w}_{GS}(j)].$$

For $0 < \alpha < 1$, one has an approximate confidence interval for γ of level $1 - \alpha$ with lower bound

$$\gamma_{GSL\alpha} = g - t_{k-1,\alpha} \hat{\sigma}_{GS}(g)$$

and upper bound

$$\gamma_{GSU\alpha} = g + t_{k-1,\alpha} \hat{\sigma}_{GS}(g).$$

Even in the elementary case of g defined as in (7) and γ defined by (6), the variance estimate $\hat{\sigma}^2_{GS}(g)$ is not the same as $\hat{\sigma}^2(g)$. Nonetheless, it is not difficult to verify that $\hat{\sigma}^2_{GS}(g)$ has expectation $\sigma^2(g)$ if the n_h/k are integers for each sample from population h. In addition, if the X_{ih} are normally distributed, then $(k-1)\hat{\sigma}^2_{GS}(g)/\sigma^2(g)$ has a chi-squared distribution on $k - 1$ degrees of freedom, and $(g - \gamma)/\hat{\sigma}_{GS}(g)$ has a t distribution on $k - 1$ degrees of freedom. Thus the probability is exactly $1 - \alpha$ that $\gamma_{GSL\alpha} \le \gamma \le \gamma_{GSU\alpha}$.

In more complex applications under study, large-sample approximations are required similar to those for grouped jackknifing with simple random sampling with replacement. In addition,

computational constraints require that k not increase even if the sample sizes n_h become large. In this case, grouped jackknifing applies when independent random variables Y_{ih}, $1 \le i \le n_h$, $1 \le h \le H$, are available such that, for each population h, the Y_{ih} are identically distributed with common mean 0 and common finite variance $\upsilon_h^2 > 0$. Let

$$\bar{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} Y_{ih},$$

$$\bar{Y}_h[\mathbf{w}] = \{n_h[\mathbf{w}]\}^{-1} \sum_{i=1}^{n_h} w_{ih} Y_{ih},$$

$$f = \sum_{h=1}^{H} \bar{Y}_h,$$

and

$$f[\mathbf{w}] = \sum_{h=1}^{H} \bar{Y}_h[\mathbf{w}].$$

The Y_{ih} are selected so that g is well approximated by $\gamma + f$ and $g[\mathbf{w}]$ is well approximated by $\gamma + f[\mathbf{w}]$. The approximation errors

$$R_{GS} = g - \gamma - f \tag{9}$$

and

$$R_{GS}[\mathbf{w}_{GS}(j)] = g[\mathbf{w}_{GS}(j)] - \gamma - f[\mathbf{w}_{GS}(j)] \tag{10}$$

must be small for large sample sizes n_h, $1 \le h \le H$. To be more specific, it is assumed that

$$E(R_{GS}^2)/\sigma^2(f) \to 0 \tag{11}$$

and

$$\max_{1 \le j \le k} E(\{R_{GS} - R_{GS}[\mathbf{w}_{GS}(j)]\}^2)/\sigma^2(f) \to 0 \tag{12}$$

as the sample sizes n_h increase for all populations h. For g defined by (7) and γ defined by (6), these conditions hold trivially for $Y_{ih} = c_h(X_{ih} - \mu_h)$ and $\upsilon_h^2 = c_h^2 \tau_h^2$. Under these conditions, the variance ration $\sigma^2(g)/\sigma^2(f)$ converges to 1 as the sample sizes n_h all become large, and the bias $E(g) - \gamma$ is sufficiently small so that $[E(g) - \gamma]/\sigma(g)$ converges to 0 as the sample sizes n_h increase. In addition, for large sample sizes n_h, $(k-1)\hat{\sigma}_{GS}^2(g)/\sigma^2(g)$ has an approximate chi-square distribution on $k-1$ degrees of freedom, and $(g-\gamma)/\hat{\sigma}_{GS}(g)$ has an approximate t distribution on $k-1$ degrees of freedom. Thus, as the n_h all become large, the probability approaches $1-\alpha$ that $\gamma_{GSL\alpha} \le \gamma \le \gamma_{GSU\alpha}$.

In section 2, sample h corresponds to examinees in Administration h of an educational test. In the specific example presented, only $H = 2$ administrations are examined. Equating is studied, so the parameter γ in the application under study may be the score on Administration 1 of the test that may be regarded as equivalent to a specified score s on Administration 2. The parameter γ can be regarded as the equating result that would be obtained were the population distribution of examinee responses known for each administration.

1.5 Jackknifing Comparisons

In applications in this report to linking of forms, a major issue involves comparison of different linking functions based on different sets of anchor items. The basic analysis is readily accomplished given the sampling procedure and grouping procedure in section 1.4. For some integer $D > 1$, consider M different estimates g_m, $1 \le m \le M$, for the respective parameters γ_m. In typical applications in this report, m will correspond to a particular set of anchor items that might be employed in equating and g_m will be the estimate for anchor set m of a specific equating result γ_m that would be obtained were all data available on all population members. For example, for a specific raw score point, there may be M different raw-to-raw conversions g_m, $1 \le m \le M$, from one form to another that have been produced from equating with M different sets of anchor items. Of interest here is the variability of the parameters γ_m for $1 \le m \le M$. It is assumed that the g_m have finite variances. Let the average of the γ_m, $1 \le m \le M$, be

$$\gamma. = M^{-1} \sum_{m=1}^{M} \gamma_m. \tag{13}$$

One simple measure of the variability of the parameters γ_m, $1 \le m \le M$, is their sample variance

$$\sigma_\gamma^2 = (M-1)^{-1} \sum_{m=1}^{M} (\gamma_m - \gamma.)^2. \tag{14}$$

One may estimate the average parameter value $\gamma.$ by the corresponding average estimate

$$g. = M^{-1} \sum_{m=1}^{M} g_m, \tag{15}$$

and the sample variance σ_γ^2 of the parameters γ_m, $1 \le m \le M$, may be estimated by the corresponding sample variance

$$\hat{\sigma}_\gamma^2 = (M-1)^{-1} \sum_{m=1}^{M} (g_m - g.)^2 \tag{16}$$

of the estimates g_m, $1 \le m \le M$. The average $g.$ has expectation

$$E(g.) = M^{-1} \sum_{m=1}^{M} \gamma_m. \tag{17}$$

If one recalls that $E(Y^2) = [E(Y)]^2 + \sigma^2(Y)$ if Y is a random variable with a finite variance, then one finds that the expectation of the sample variance $\hat{\sigma}_\gamma^2$ is

$$E(\hat{\sigma}_\gamma^2) = (M-1)^{-1} \sum_{m=1}^{M} [E(g_m) - E(g.)]^2 + (M-1)^{-1} \sum_{m=1}^{M} \sigma^2(g_m - g.). \tag{18}$$

In typical cases, the variance estimate $\hat{\sigma}_\gamma^2$ has a positive bias as an estimate of the sample variance σ_γ^2 of the γ_m. This condition is readily observed in the elementary case in which one has independent random vectors $\mathbf{X}_{ih}$ with mean $\boldsymbol{\mu}_h$ and positive-definite covariance matrix $\mathbf{C}_h > 0$ for $1 \le i \le n_h$ and $1 \le h \le H$. Let coordinate m of $\mathbf{X}_{ih}$ be X_{mih}, and let coordinate m of $\boldsymbol{\mu}$ be μ_m. For $1 \le m \le H$, consider estimation of a linear combination

$$\gamma_m = \sum_{h=1}^{H} c_{mh} \mu_h \tag{19}$$

of the means μ_{mh}, $1 \le h \le H$, for some real numbers c_{mh}, $1 \le h \le H$. For example, if $c_{mh} = H^{-1}$ for each population h, then γ_m is the average $\bar{\mu}_m$ of the population means μ_{mh}. The conventional estimate of γ_m is the linear combination

$$g_m = \sum_{h=1}^{H} c_{mh} \bar{X}_{mh} \tag{20}$$

of the sample means

$$\bar{X}_{mh} = n_h^{-1} \sum_{i=1}^{n_h} X_{mih}.$$

The mean of g_m is γ_m, so that g_m is unbiased, and the mean of $g.$ is $\gamma.$. For a vector $\mathbf{b}$, let $\mathbf{b}'$ denote its transpose. Then the expectation of $\hat{\sigma}_\gamma^2$ is

$$E(\hat{\sigma}_\gamma^2) = \sigma_\gamma^2 + \sum_{m=1}^{M} \sigma^2(g_m - g.),$$

where the variance of $g_m - g.$ is

$$\sigma^2(g_m - g.) = \sum_{h=1}^{H} \mathbf{b}'_{mh} \mathbf{C}_h \mathbf{b}_{mh} / n_h$$

and $\mathbf{b}_{mh}$ is the M-dimensional vector with coordinate $b_{m'mh}$, $1 \le m' \le M$, such that $b_{m'mh}$ is $c_{mh}(M-1)/M$ for $m' = d$ and $b_{m'mh}$ is $-c_{m'h}/M$ for $m' \ne d$.

To investigate this bias in estimation of the sample variance σ_γ^2 of the parameters γ_m, $1 \le m \le M$, grouped jackknifing for stratified random sampling may be employed. The basic requirement is that each g_m satisfy the requirements for grouped jackknifing described in section 1.4 for the case of stratified random sampling. Consider the following conditions. Let $\mathbf{0}$ be the M-dimensional vector with all coordinates 0. Let $\mathbf{Y}_{ih}$, $1 \le i \le n_h$, $1 \le h \le H$, be independent M-dimensional vectors with coordinates Y_{mih} , $1 \le m \le M$, such that, for sample h, the $\mathbf{Y}_{ih}$ are identically distributed with common mean $\mathbf{0}$ and common positive-definite covariance matrix $\boldsymbol{\Upsilon}_h$. Let row m and column m' of $\boldsymbol{\Upsilon}_h$ be $\Upsilon_{mm'h}$. Let

$$\bar{Y}_{mh} = n_h^{-1} \sum_{i=1}^{n_h} Y_{mhi}, \tag{21}$$

let the average of the Y_{imh} over m be

$$Y_{\cdot ih} = M^{-1} \sum_{m=1}^{M} Y_{mih}, \tag{22}$$

and let

$$\bar{Y}_{\cdot h} = n_h^{-1} \sum_{i=1}^{n_h} Y_{\cdot hi}, \tag{23}$$

so that

$$\bar{Y}_{\cdot h} = M^{-1} \sum_{m=1}^{M} \bar{Y}_{mh}.$$

Similarly, for the weight function $\mathbf{w}$ with integer coordinates $w_{ih} \ge 0$, $1 \le i \le n_h$, $1 \le h \le H$, let

$$\bar{Y}_{mh}[\mathbf{w}] = \{n_h[\mathbf{w}]\}^{-1} \sum_{i=1}^{n_h} w_{ih} Y_{mih} \tag{24}$$

and

$$\bar{Y}_{\cdot h}[\mathbf{w}] = \{n_h[\mathbf{w}]\}^{-1} \sum_{i=1}^{n_h} w_{ih} Y_{\cdot ih} \tag{25}$$

whenever $n_h[\mathbf{w}] > 0$. Let

$$f_m = \sum_{h=1}^{H} \bar{Y}_{mh}, \tag{26}$$

and let

$$f_m[\mathbf{w}] = \sum_{h=1}^{H} \bar{Y}_{mh}[\mathbf{w}]. \tag{27}$$

Note that f_m has variance

$$\sigma^2(f_m) = \sum_{h=1}^{H} \Upsilon_{mmh}/n_h.$$

The Y_{mih} are selected so that g_m is well approximated by $\gamma_m + f_m$ and $g_m[\mathbf{w}]$ is well approximated by $\gamma_m + f_m[\mathbf{w}]$. The approximation errors

$$R_{mGS} = g_m - \gamma_m - f_m \tag{28}$$

and

$$R_{mGS}[\mathbf{w}_{GS}(j)] = g_m[\mathbf{w}_{GS}(j)] - \gamma_m - f_m[\mathbf{w}_{GS}(j)] \tag{29}$$

must be small for large sample sizes n_h, $1 \leq h \leq H$. To be more specific, it is assumed that

$$E(R_{mGS}^2)/\sigma^2(f_m) \to 0 \tag{30}$$

and

$$\max_{1 \leq j \leq k} E(\{R_{mGS} - R_{mGS}[\mathbf{w}_{GS}(j)]\}^2)/\sigma^2(f_m) \to 0 \tag{31}$$

as the sample sizes n_h increase for all populations h.

Because the matrices Υ_m are positive definite for $1 \leq m \leq M$, (30) and (31) imply that (11) and (12) hold whenever c_m, $1 \leq m \leq M$, are real numbers, some c_m is not 0,

$$g = \sum_{m=1}^{M} c_m g_m,$$

$$\gamma = \sum_{m=1}^{M} c_m \gamma_m,$$

and

$$Y_{ih} = \sum_{m=1}^{M} c_m Y_{mih}.$$

It follows that $\sigma^2(g)/\sigma^2(f)$ converges to 1, the bias $E(g) - \gamma$ is sufficiently small that $[E(g) - \gamma]/\sigma(g)$ approaches 0 as the sample sizes n_h all become large, and the ratio $(k-1)\hat{\sigma}_{GS}^2(g)/\sigma^2(g)$ has an approximate chi-square distribution on $k-1$ degrees of freedom. Consideration of the differences $g_m - g.$ shows that the bias

$$\Delta = E(\hat{\sigma}_\gamma^2) - \sigma_\gamma^2 \tag{32}$$

is well approximated by

$$\Delta_0 = (M-1)^{-1} \sum_{m=1}^{M} \sigma^2(g_m - g.) \tag{33}$$

in the sense that Δ/Δ_0 converges to 1 as the sample sizes n_h all become large. If $f.$ is the average $M^{-1}\sum_{m=1}^{M} f_m$ and if

$$\Delta_1 = (M-1)^{-1}\sum_{m=1}^{M}\sigma^2(f_m - f.), \tag{34}$$

then Δ/Δ_1 also converges to 1 as the sample sizes n_h all become large.

One may approximate the bias Δ with the grouped jackknife by use of

$$\hat{\Delta}_{GS} = (M-1)^{-1}\sum_{m=1}^{M}\hat{\sigma}^2_{GS}(g_m - g.), \tag{35}$$

so that a bias-corrected estimate of σ^2_γ is

$$\hat{\sigma}^2_{GS\gamma} = \hat{\sigma}^2_\gamma - \hat{\Delta}_{GS}. \tag{36}$$

To understand this correction, consider a large-sample approximation in which the fraction of observations from each population has a positive limit. For this purpose, let $n_+ = \sum_{h=1}^{H} n_h$ be the total sample size. Let n_+ become large and, for each population h, let the ratio n_h/n_+ approach a positive constant ω_h. Then the large-sample distribution of $\hat{\Delta}_{GS}/\Delta$ may be studied by use of general results concerning the distribution of quadratic functions of multivariate normal random variables (Box, 1954). Let $\mathbf{Q}$ be the M by M matrix with row m and column m' equal to $1 - M^{-1}$ if $m = m'$ and equal to $-M^{-1}$ if $m \neq m'$. Let

$$\mathbf{\Omega} = \mathbf{Q}\sum_{h=1}^{H}\omega_h \mathbf{\Upsilon}_h. \tag{37}$$

Let tr denote a trace of a square matrix. Then $\hat{\Delta}_{GS}/\Delta$ converges in distribution to a positive random variable Z with expectation 1 and with variance

$$E(Z) = 2(k-1)^{-1}\frac{\mathrm{tr}(\mathbf{\Omega}\mathbf{\Omega})}{[\mathrm{tr}(\mathbf{\Omega})]^2}.$$

The trace $\mathrm{tr}(\mathbf{\Omega}\mathbf{\Omega})$ is the sum of the squares of the $M-1$ nonzero eigenvalues of $\mathbf{\Omega}$, while $\mathrm{tr}(\mathbf{\Omega})$ is the sum of the $M-1$ nonzero eigenvalues of $\mathbf{\Omega}$ (Box, 1954). The Cauchy-Schwarz inequality may be used to demonstrate that Z has variance less than $2/(k-1)$ but at least as large as $2/[(k-1)(M-1)]$. Note that $|\Delta$ is well approximated by Δ_1, and Δ_1 is of order of magnitude equal to the largest of the inverse sample sizes n_h^{-1} for $1 \leq h \leq H$. Thus the bias Δ is small in large samples, and the bias correction $\hat{\Delta}_{GS}$ is also small in such cases.

An exact result is available in the special case of $g_m = \gamma_m + f_m$ for $1 \leq m \leq M$, γ_m constant over m, n_h/k an integer for $1 \leq h \leq H$, and Y_{mhi} independent normal random variables with common variance for all m and i. Application of standard results from two-way analysis of variance with one observation per cell shows that $\sigma_\gamma^2 = 0$, the bias Δ is $\sigma^2(f_1)$, and $(k-1)(M-1)\hat{\Delta}_{GS}/\Delta$ has a chi-squared distribution on $(k-1)(M-1)$ degrees of freedom, so that $\hat{\Delta}_{GS}/\Delta$ has mean 1 and variance $2/[(k-1)(M-1)]$. The variance of $\hat{\Delta}_{GS}/\Delta$ decreases as the number k of groups increases and as the number M of estimates g_m, $1 \leq m \leq M$, increases. In addition, $\sigma^2(f_1)$ is approximately proportional to the total sample size n_+^{-1}, and the ratio $\hat{\sigma}_\gamma^2/\hat{\Delta}_{GS}$ has an F distribution with $M-1$ and $(k-1)(M-1)$ degrees of freedom. Note that in this case, there is a substantial probability that bias-corrected estimate $\hat{\sigma}^2_{GS\gamma}$ is negative, even though the estimated quantity σ_γ^2 must be nonnegative.

1.6 Randomly Selected Estimates

The analysis in section 1.5 raises a rather basic issue in the context of equating. In many cases, it is assumed implicitly in equating that different selections of anchor sets should lead to the same basic equating results. For example, except for sampling error, conversions of scores on a new form to an old form should be the same. In practice, errors are encountered both due to the failure of equating assumptions and due to sampling error. One simple assessment considers a randomly selected estimate g_S, where S is uniformly distributed on the integers 1 to M and independent of the g_m. Thus g_S is g_m with probability 1. The estimate g_S reflects results of equating if the anchor set really is randomly selected. The expected value of g_S is the average

$$E(g_S) = E(g.) = M^{-1} \sum_{m=1}^{M} g_m \tag{38}$$

of the expectations $E(g_m)$, $1 \leq m \leq M$. The variance $\sigma^2(g_S)$ of g_S has two components, the expected conditional variance of g_S given S and the variance of the expected conditional mean of g_S given S (Rao, 1973, p. 97). It follows that

$$\sigma^2(g_S) = \frac{M-1}{M}\sigma_\gamma^2 + M^{-1} \sum_{m=1}^{M} \sigma^2(g_m). \tag{39}$$

The bias $E(g_S) - \gamma.$ is sufficiently small that

$$\frac{E(g_S) - \gamma.}{\left[M^{-1} \sum_{m=1}^{M} \sigma^2(g_m)\right]^{1/2}} \to 0$$

as the sample sizes n_h all become large.

In addition, the random difference $g_S - g.$ and the average estimate $g.$ are uncorrelated. To verify this claim, note that (38) implies that $g_S - g.$ has expectation 0. Thus the covariance of $g_S - g.$ and $g.$ is the expectation

$$E([g_S - g.]g.) = M^{-1} \sum_{m=1}^{M} E([g_m - g.]g.) = E([g. - g.]g.) = 0. \tag{40}$$

As in (39),

$$\sigma^2(g_S - g.) = \frac{M-1}{M}\sigma_\gamma^2 + M^{-1} \sum_{m=1}^{M} \sigma^2(g_m - g.). \tag{41}$$

Combination of (39), (40), and (41) leads to

$$\sigma^2(g_S) = \frac{M-1}{M}\sigma_\gamma^2 + \sigma^2(g.) + M^{-1} \sum_{m=1}^{M} \sigma^2(g_m - g.). \tag{42}$$

With jackknifing, (35) implies that $\sigma^2(g_S)$ may be estimated by

$$\hat{\sigma}_{GS}^2(g_S) = \frac{M-1}{M}\hat{\sigma}_\gamma^2 + \hat{\sigma}_{GS}^2(g.). \tag{43}$$

In (43), the first component on the right-hand side assesses variability in the estimates g_m, $1 \le m \le M$, and the second component measures the variability of the average estimate $g.$. As the sample sizes n_h increase for all populations h, $\sigma^2(g_S)$ approaches $[(M-1)/M]\sigma_\gamma^2$. If the γ_m are not all the same, then $\sigma_\gamma^2 > 0$ and this limiting variance is positive. No matter how large are the samples, accuracy is then limited by the inconsistency of parameters γ_m, $1 \le m \le M$, for different anchor choices. Interpretation is to some degree made more complicated because anchor items are not really chosen at random. Nonetheless, the analysis can provide some measure of the impact of anchor choice.

1.7 Overlapping Anchor Sets

In many common cases, including the examples to be presented in section 3, one anchor set was actually employed in equating. For instance, in one case, an anchor set consisted of 28 items. Alternate anchor sets are obtained by deletion of single items of groups of items. Thus the possible anchor sets used are very similar. For example, one might consider 28 anchor sets derived from the original 28 items by deletion of one item. One would expect that this similarity of anchor sets would result in less variability related to choice of anchor sets than would be encountered

were completely different anchor items employed. It is possible to try to estimate the effects of more thorough changes in anchor sets from the limited selections available, but some reasonable assumptions must be made. These assumptions can be similar to those used in jackknifing. They are relevant when anchor items or anchor blocks are selected at random from a large enough finite population so that corrections for finite populations can be ignored. The extent to which this model is realistic can be debated, for anchor items are not chosen at random in typical cases. Nonetheless, the analysis may still provide insight into reasonable expectations for variability. Let there be M anchor items (or anchor blocks) I_m, $1 \leq m \leq M$, selected at random and used in an assessment.

For a specific choice of anchor items I_m, $1 \leq m \leq M$, let $g_{\mathbf{I}m}$, $1 \leq m \leq M$, represent an equating result based on use of the anchor items $I_{m'}$ for $m' \neq m$, and let $g_{\mathbf{I}0}$ be an equating result based on use of all the anchor items I_m, $1 \leq m \leq M$, and let $g_{\mathbf{I}m}$ estimate $\gamma_{\mathbf{I}m}$. Let

$$\gamma_{\mathbf{I}\cdot} = M^{-1} \sum_{m=1}^{M} \gamma_{\mathbf{I}m}. \tag{44}$$

and let

$$\sigma^2_{\mathbf{I}\gamma} = (M-1)^{-1} \sum_{m=1}^{M} (\gamma_{\mathbf{I}m} - \gamma_{\mathbf{I}\cdot})^2. \tag{45}$$

Define the estimates

$$g_{\mathbf{I}\cdot} = M^{-1} \sum_{m=1}^{M} g_{\mathbf{I}m} \tag{46}$$

and

$$\hat{\sigma}^2_{\mathbf{I}\gamma} = (M-1)^{-1} \sum_{m=1}^{M} (g_{\mathbf{I}m} - g_{\mathbf{I}\cdot})^2. \tag{47}$$

Assume that each $g_{\mathbf{I}m}$ has a finite mean and a finite variance.

To treat randomly selected estimates, for $0 \leq m \leq M$, let g_m be the random estimate with value $g_{\mathbf{I}m}$ if anchor items $I_{m'}$, $1 \leq m' \leq M$, are selected. Similarly, let γ_m be the random variable with value $\gamma_{\mathbf{I}m}$ if the anchor items $I_{m'}$, $1 \leq m' \leq M$, are selected. Let $\gamma_{\cdot}$ denote the random variable with value $\gamma_{\mathbf{I}\cdot}$ if I_m, $1 \leq m \leq M$, is selected, and let σ^2_γ denote the random variable with value $\sigma^2_{\mathbf{I}\gamma}$ if I_m, $1 \leq m \leq M$, is selected. The estimated equating result in practice is g_0. The estimate g_0 in effect estimates the expectation $E(\gamma_0)$ of γ_0. The variance of $\sigma^2(g_0)$ is the sum of two components. The first component is the expected value $\sigma^2_1(g_0)$ of the random variable $\sigma^2(g_0)$ with value equal to $\sigma^2(g_{\mathbf{I}0})$ if $I_{m'}$, $1 \leq m' \leq M$, is selected. The second component is the variance

$\sigma_2^2(g_0)$ of the random variable with value $E(g_{\mathbf{I}0})$ if $I_{m'}$, $1 \le m' \le M$, is selected (Rao, 1973, p. 97). In this section, conditions are developed under which both components can be approximated. The first set of conditions permits use of jackknifing to approximate $\sigma^2(g_{\mathbf{I}m})$ for any possible selection of anchor items $I_{m'}$, $1 \le m' \le M$. This set of conditions is essentially the same as in section 1.5. Additional conditions are then imposed to permit approximation of $\sigma_2^2(g_0)$. These conditions are somewhat related to those developed in section 1 for jackknifing for simple random sampling, but they apply to items rather than to examinees.

It is assumed that, for any selection of anchor items $I_{m'}$, $1 \le m' \le M$, the $g_{\mathbf{I}m}$ satisfy the basic conditions for grouped jackknifing in stratified random samples that were described in section 1.5. Thus one has independent pairs $(Y_{\mathbf{I}0ih}, \mathbf{Y}_{\mathbf{I}ih})$, $1 \le i \le n_h$, $1 \le h \le H$, where $\mathbf{Y}_{\mathbf{I}ih}$ has coordinates $Y_{\mathbf{I}mih}$, $1 \le m \le M$. For population h, the pairs $(Y_{\mathbf{I}0ih}, \mathbf{Y}_{\mathbf{I}ih})$ are identically distributed for $1 \le i \le n_h$, $Y_{\mathbf{I}0ih}$ has mean 0 and finite and positive variance $\Upsilon_{\mathbf{I}00h}$, and $\mathbf{Y}_{\mathbf{I}ih}$ has mean $\mathbf{0}$ and finite positive-definite covariance matrix $\boldsymbol{\Upsilon}_{\mathbf{I}h}$. For $0 \le m \le M$, define

$$\bar{Y}_{\mathbf{I}mh} = n_h^{-1} \sum_{i=1}^{n_h} Y_{\mathbf{I}mhi}, \tag{48}$$

let

$$\bar{Y}_{\mathbf{I}\cdot h} = M^{-1} \sum_{m=1}^{M} \bar{Y}_{\mathbf{I}mh}, \tag{49}$$

and let

$$\bar{Y}_{\mathbf{I}mh}[\mathbf{w}] = \{n_h[\mathbf{w}]\}^{-1} \sum_{i=1}^{n_h} w_{ih} Y_{\mathbf{I}mih} \tag{50}$$

whenever $n_h[\mathbf{w}] > 0$. Let

$$f_{\mathbf{I}m} = \sum_{h=1}^{H} \bar{Y}_{\mathbf{I}mh}, \tag{51}$$

and let

$$f_{\mathbf{I}m}[\mathbf{w}] = \sum_{h=1}^{H} \bar{Y}_{\mathbf{I}mh}[\mathbf{w}]. \tag{52}$$

Let

$$f_{\mathbf{I}\cdot} = \sum_{h=1}^{H} \bar{Y}_{\mathbf{I}\cdot h}, \tag{53}$$

Let the approximation errors

$$R_{\mathbf{I}mGS} = g_{\mathbf{I}m} - \gamma_{\mathbf{I}m} - f_{\mathbf{I}m} \tag{54}$$

and

$$R_{\mathbf{I}mGS}[\mathbf{w}_{GS}(j)] = g_{\mathbf{I}m}[\mathbf{w}_{GS}(j)] - \gamma_{\mathbf{I}m} - f_{\mathbf{I}m}[\mathbf{w}_{GS}(j)] \tag{55}$$

satisfy the conditions that

$$E(R^2_{\mathbf{I}mGS})/\sigma^2(f_{\mathbf{I}m}) \to 0 \tag{56}$$

and

$$\max_{1\le j\le k} E(\{R_{\mathbf{I}mGS} - R_{\mathbf{I}mGS}[\mathbf{w}_{GS}(j)]\}^2)/\sigma^2(f_{\mathbf{I}m}) \to 0 \tag{57}$$

as the sample sizes n_h increase for all populations h.

Under these conditions, as the sample sizes n_h approach ∞, the following limiting relationships hold:

$$\frac{\sigma^2(g_{\mathbf{I}0})}{\sigma^2(f_{\mathbf{I}0})} \to 1, \tag{58}$$

$$\frac{E(g_{\mathbf{I}0}) - \gamma_{\mathbf{I}0}}{\sigma(g_{\mathbf{I}0})} \to 0, \tag{59}$$

and the ratio $(k-1)\hat{\sigma}^2_{GS}(g_{\mathbf{I}0})/\sigma^2(g_{\mathbf{I}0})$ converges in distribution to a random variable with a chi-square distribution on $k-1$ degrees of freedom. If $\hat{\sigma}^2_{GS}(g_0)$ denotes the random variable with value $\hat{\sigma}^2_{GS}(g_{\mathbf{I}0})$ if the anchor items I_m, $1 \le m \le M$, are selected, then one can certainly approximate $\sigma^2_1(g_0)$ by use of $\hat{\sigma}^2_{GS}(g_0)$.

To estimate $\sigma^2_2(g_0)$ requires some assumptions concerning the parameters $\gamma_{\mathbf{I}m}$ and the random variables $Y_{\mathbf{I}mih}$ for $0 \le m \le M$. The assumption made here is that the parameter $\gamma_{\mathbf{I}m}$ has a decomposition

$$\gamma_{\mathbf{I}m} = \begin{cases} \beta + (M-1)^{-1}\sum_{m'\ne m}\nu(I_{m'}) + \zeta_{\mathbf{I}m}, & 1 \le m \le M, \\ \beta + M^{-1}\sum_{m'=1}^{M}\sum_{m'=1}^{M}\nu(I_{m'}) + \zeta_{\mathbf{I}0}, & m = 0, \end{cases} \tag{60}$$

where the constants $\zeta_{\mathbf{I}m}$ are remainder terms, and the random variable $Y_{\mathbf{I}mih}$ has the decomposition

$$Y_{\mathbf{I}mih} = \begin{cases} Z_{ih} + (M-1)^{-1}\sum_{m'\ne m}U_{ih}(I_{m'}) + e_{\mathbf{I}m}, & 1 \le m \le D, \\ Z_{i0} + M^{-1}\sum_{m'=1}^{M}\sum_{m'=1}^{M}\nu(I_{m'}) + e_{\mathbf{I}mih}, & m = 0. \end{cases} \tag{61}$$

where the random variables $e_{\mathbf{I}mih}$ are remainder terms. In (60), 0 is the average of the $\nu(A)$ over all possible anchor items A, and σ^2_ν is the variance of a random variable ν_m with value $\nu(I_m)$ if $\mathbf{I}$ is randomly selected. In (61), the components Z_{ih} and $U_{ih}(I_{m'})$ are all independently distributed. For each sample h, the Z_{ih} are identically distributed with mean 0 and finite variance σ^2_{Zh} and the $U_{ih}(I_{m'})$ are identically distributed for each anchor item I_m and have mean 0 and finite variance $\sigma^2_{Uh}(I_{m'}) > 0$. The parameter σ^2_{Um} denotes the mean of the random variable with value $\sigma^2_{Uh}(I_{m'})$ if the $I_{m'}$ are selected at random for $1 \le m' \le M$.

In (60),

$$\gamma_{\mathbf{I}\cdot} = \beta + M^{-1}\sum_{m=1}^{M}\nu(I_m) + \zeta_{\mathbf{I}\cdot},$$

where

$$\zeta_{\mathbf{I}\cdot} = M^{-1}\sum_{m=1}^{M}\zeta_{\mathbf{I}m}.$$

In (61),

$$Y_{\mathbf{I}\cdot ih} = M^{-1}\sum_{m=1}^{M} Y_{\mathbf{I}mih} = Z_{ih} + M^{-1}\sum_{m=1}^{M} U_{ih}(I_m) + e_{\mathbf{I}\cdot ih},$$

where

$$e_{\mathbf{I}m\cdot ih} = M^{-1}\sum_{m=1}^{M} e_{\mathbf{I}mih}.$$

Thus

$$\delta_{\mathbf{I}} = \gamma_{\mathbf{I}0} - \gamma_{\mathbf{I}\cdot} = \zeta_{\mathbf{I}0} - \zeta_{\mathbf{I}\cdot}$$

and

$$V_{\mathbf{I}ih} = Y_{\mathbf{I}0ih} - Y_{\mathbf{I}\cdot ih} = e_{\mathbf{I}0ih} - e_{\mathbf{I}\cdot ih}.$$

Let $\zeta.$ be the random variable with value $\zeta_{\mathbf{I}\cdot}$ if $I_{m'}$, $1 \le m' \le M$, is selected, let W_{Yh} be the random variable with value $E(e^2_{\mathbf{I}\cdot ih})$ if $I_{m'}$, $1 \le m' \le M$, is selected, let δ be the random variable with value $\delta_{\mathbf{I}}$ if $I_{m'}$, $1 \le m' \le M$, is selected, and let W_{Vh} be the random variable with value $E(V^2_{\mathbf{I}ih})$ if $I_{m'}$, $1 \le m' \le M$, is selected. The approximate methods used in this section require that $\zeta.$, δ, $E(W_{Yh})$, and $E(W_{Vh})$ all be small relative to σ^2_ν/M. To examine this claim, consider the simplified case in which $\zeta_{\mathbf{I}\cdot}$, $\zeta_{\mathbf{I}0}$, $e_{\mathbf{I}\cdot ih}$, and $e_{\mathbf{I}0ih}$ are all 0 for any anchor items $I_{m'}$, $1 \le m' \le M$. In this case, comparison with delete-1 jackknifing for sample means shows that $\sigma^2_2(g_0)$ is σ^2_ν/M and

$$\sigma^2_\gamma = (M-1)^{-3}\sum_{m=1}^{M}(\nu_m - \nu.)^2,$$

where $\nu.$ is the average $M^{-1}\sum_{m=1}^{M}\nu_m$ of the ν_m, $1 \le m \le M$. The expectation of σ^2_γ is then $(M-1)^{-2}\sigma^2_\nu$, so that $M^{-1}(M-1)^2\sigma^2_\gamma$ has expectation $\sigma^2_2(g_0)$. In addition, g_0 and $g.$ are equal. If $\hat{\sigma}^2_{GS}(g.)$ denotes the random variable with value $\hat{\sigma}^2_{GS}(g_{\mathbf{I}\cdot})$ if the anchor items I_m, $1 \le m \le M$, are selected, then $\sigma^2_1(g_0)$ may be approximated by $\hat{\sigma}^2_{GS}(g.)$ as well as by $\hat{\sigma}^2_{GS}(g_0)$. If

$$\hat{\Delta}_{\mathbf{I}GS} = (M-1)^{-1}\sum_{m=1}^{M}\hat{\sigma}^2_{GS}(g_{\mathbf{I}m} - g_{\mathbf{I}\cdot}) \tag{62}$$

for anchor items $I_{m'}$, $1 \le m' \le M$, and if $\hat{\Delta}_{GS}$ is the random estimate with value $\hat{\Delta}_{\mathbf{I}GS}$ if the $I_{m'}$, $1 \le m' \le M$, are selected, then $\sigma_2^2(g_0)$ may be approximated by $(M-1)^2[\hat{\sigma}_\gamma^2 - \hat{\Delta}_{GS}]$. Approximations are most satisfactory if the number k of groups and the number M of anchor items is large.

If σ_ν^2 is 0 and if, for each population h and the $U_{ih}(I_{m'})$ all have the same variance, then $\sigma_\gamma^2 = 0$ and

$$F_{GS} = (M-1)^2 \hat{\sigma}_\gamma^2 / \hat{\Delta}_{GS} \tag{63}$$

has an approximate F distribution on $M-1$ and $(M-1)(k-1)$ degrees of freedom. The result is exact if each ratio n_h/k is an integer and if the $U_{ih}(I_{m'})$ have normal distributions.

The suggested estimate of $\sigma_{GS}^2(g_0)$ based on the case of no remainder errors is

$$\hat{\sigma}_{GS}^2(g_0) = \hat{\sigma}_{GS}^2(g_0) + M^{-1}(M-1)^2[\hat{\sigma}_\gamma^2 - \hat{\Delta}_{GS}]. \tag{64}$$

A slight modification $\bar{\sigma}_{GS}^2(g_0)$ of this estimate has the attraction that it can be computed from a two-way array of estimates $g_m[\mathbf{w}_{GS}(j)]$, $1 \le j \le k$, $1 \le m \le M$. Let $g_m[\mathbf{w}_{GS}(j)]$ be the random estimate with value $g_{\mathbf{I}m}[\mathbf{w}_{GS}(j)]$ if items $I_{m'}$, $1 \le m' \le M$, are selected. The average of the $g_m[\mathbf{w}_{GS}(j)]$, $1 \le m \le M$, can be denoted by $\bar{g}.[\mathbf{w}_{GS}(j)]$, and the average of the $g.[\mathbf{w}_{GS}(j)]$, $1 \le j \le k$, can be denoted by $g..$ This modified estimate $\bar{\sigma}_{GS}^2(g_0)$ is the same as $\hat{\sigma}_{GS}^2(g_0)$ if $\zeta_{\mathbf{I}\cdot}$, $\zeta_{\mathbf{I}0}$, $e_{\mathbf{I}\cdot ih}$, and $e_{\mathbf{I}0ih}$ are all 0. To define the modified estimate, let $g_m[\mathbf{w}_{GS}(j)]$ be the random estimate with value $g_{\mathbf{I}m}[\mathbf{w}_{GS}(j)]$ if items $I_{m'}$, $1 \le m' \le M$, are selected. The average of the $g_m[\mathbf{w}_{GS}(j)]$, $1 \le m \le M$, can be denoted by $g.[\mathbf{w}_{GS}(j)]$, and the average of the $g.[\mathbf{w}_{GS}(j)]$, $1 \le j \le k$, can be denoted by $\bar{g}..$ For each item I_m, the average of the $g_m[\mathbf{w}_{GS}(j)]$, $1 \le j \le k$, may be denoted by $\bar{g}_m$. One has

$$\bar{\sigma}_{GS}^2(g_0) = \frac{k-1}{k} \sum_{j=1}^{k} \{g.[\mathbf{w}_{GS}(j)] - \bar{g}.\}^2,$$

$$\bar{\sigma}_\gamma^2 = (M-1)^{-1} \sum_{m=1}^{M} (\bar{g}_m - \bar{g}.)^2,$$

and

$$\bar{\Delta}_{GS} = \frac{k-1}{k}(M-1)^{-1} \sum_{m=1}^{M} \{g_m[\mathbf{w}_{GS}(j)] - g.[\mathbf{w}_{GS}(j)] - \bar{g}_m + \bar{g}.\}^2.$$

It follows that

$$\bar{\sigma}_{GS}^2(g_0) = \bar{\sigma}_{GS}^2(g_0) + M^{-1}(M-1)^2[\bar{\sigma}_\gamma^2 - \bar{\Delta}_{GS}]. \tag{65}$$

The approximations used in this section are less precise in practice than approximations used in earlier sections, for the number of anchor items is typically somewhat smaller than the number of groups and is much smaller than the actual sample sizes. Nonetheless, the basic issue remains that, as in section 1.6, the variance of g_0 does not approach 0 even for large sample sizes n_h unless the parameter γ_0 is the same for all possible anchor items I_m, $1 \leq m \leq M$.

2 IRT True-Score Equating

In the examples under study, IRT true-score equating is used for equating of two administrations, Administration 1 and Administration 2, by use of a collection of common external anchor items I_m, $1 \leq m \leq M$. The approach of Stocking and Lord (Stocking & Lord, 1983) is used with a generalized partial credit model (Muraki, 1997). This approach reflects practices of the particular testing program under study. Numerous alternatives are available (Hambleton, Swaminathan, & Rogers, 1991, ch.9). The number of examinees in Administration h is denoted by n_h. In Administration 1, each examinee receives items I_m for $M+1 \leq m \leq M_1$ where $M_1 > M+1$, and these items are used to score the examinee performance. In Administration 2, each examinee receives items I_m, $M_1+1 \leq m \leq M_2$, where $M_2 > M_1+1$, and these items are used to score the examinee. In addition, some examinees from each administration receive the common anchor items I_m, $1 \leq m \leq M$. It suffices to assume that whether an examinee in Administration h receives the external anchor items I_m, $1 \leq m \leq M$, is a random event not related to any characteristics or responses of any of the examinees.

Estimation is performed with an item-response model in which the proficiency distribution of the population of examinees for Administration 1 is a standard normal proficiency distribution, while examinees who receive Administration 2 are assumed to have a proficiency distribution that is normal with mean B and standard deviation $A > 0$. Conditional on the proficiency θ of an examinee, it is assumed that item scores for each item presented are conditionally independent. Item scores for an item I_m have possible values from 0 to $r_m - 1$, where r_m is an integer greater than 1. The conditional probability $P_m(k|\theta)$ that an examinee with proficiency θ receiving either form has response score k on a presented item I_m is assumed to satisfy the logit relationship

$$\log[P_m(k|\theta)/P_m(k-1|\theta)] = Da_m(\theta - b_m + d_{mk}),$$

where item discrimination a_m is an unknown positive real number, item difficulty b_m is an

unknown real number, and the category coefficients d_{mk}, $1 \le k \le r_m - 1$, are real numbers unknown save for the constraint that their sum is 0 (Muraki, 1997). Thus $d_{uk} = 0$ if $r_m = 2$. The constant D is fixed. It may be chosen to be 1, 1.7, or 1.702. The last choice is made here for consistency with the Parscale software used in computations at ETS.

In the case of Administration 2, the scaled examinee proficiency $\theta' = (\theta - B)/A$ has a standard normal distribution. With respect to the scaled proficiency θ', the conditional probability $P'_m(k|\theta')$ that an examinee with scaled proficiency θ' in Administration 2 has response score k on a presented item I_m satisfies the logit relationship

$$\log[P'_m(k|\theta)/P'_m(k-1|\theta)] = Da'_m(\theta' - b'_m + d'_{mk}),$$

where $a'_m = Aa_m$, $b'_m = (b_m - B)/A$, and $d'_m = d_m/A$. Marginal maximum likelihood, conditional on the items presented to each examinee, is separately employed for each Form h (Bock & Aitkin, 1981). Administration 1 yields maximum-likelihood estimates $\hat{a}_m$ for a_m, $\hat{b}_m$ for b_m, and $\hat{d}_{uk}$ for d_{uk} for $1 \le m \le m_1$ and $1 \le M_1$. Administration 2 yields maximum-likelihood estimates $\hat{a}'_m$ for a'_m, $\hat{b}'_m$ for b'_m, and $\hat{d}'_{uk}$ for m'_{uk}, $1 \le k \le r_j - 1$, for $1 \le m \le M$ and for $M_1 + 1 \le m \le M_2$.

To estimate A and B, the Stocking-Lord method is used (Stocking & Lord, 1983). Here the estimated test characteristic curves for the items used in scoring is computed for the two administrations. For Administration 1,

$$\hat{T}(\theta) = \sum_{m=1}^{M} \sum_{k=1}^{r_m-1} k\hat{P}_m(k|\theta),$$

where the estimated conditional probabilities $\hat{P}_m(k|\theta)$ are determined by the equations

$$\log[\hat{P}_m(k|\theta)/\hat{P}_m(k-1|\theta)] = D\hat{a}_m(\theta - \hat{b}_m + \hat{d}_{uk})$$

for $1 \le k \le r_m - 1$ and by the constraint that $\sum_{k=0}^{r_m-1} \hat{P}_m(k|\theta) = 1$. In like manner, for Administration 2,

$$\hat{T}'(\theta') = \sum_{m=1}^{M} \sum_{k=1}^{r_m-1} k\hat{P}'_m(k|\theta),$$

where the estimated conditional probabilities $\hat{P}'_m(k|\theta)$ are determined by the equations

$$\log[\hat{P}'_m(k|\theta)/\hat{P}'_m(k-1|\theta)] = D\hat{a}'_m(\theta - \hat{b}'_m + \hat{d}'_{mk})$$

for $1 \le k \le r_m - 1$ and by the constraint that $\sum_{k=0}^{r_m-1} \hat{P}'_m(k|\theta) = 1$. Estimates $\hat{A}$ for A and $\hat{B}$ for B are then obtained by minimizing the integral

$$\int [\hat{T}(\theta) - \hat{T}'(A\theta + B)]^2 \phi(\theta) d\theta,$$

where ϕ is the density function of the standard normal distribution. The integral must be evaluated by some numerical quadrature method. In the analysis performed in this report, the ETS convention was followed that the integral was approximated by use of 201 equally spaced quadrature points from -3 to 3.

Given $\hat{A}$ and $\hat{B}$, parameter estimates are obtained for Administration 2. Thus a_m is estimated by $\hat{a}_{u2} = \hat{a}'_m/\hat{A}$, b_m is estimated by $\hat{b}_{m2} = \hat{A}\hat{b}'_m + \hat{B}$, and d_{mk} is estimated by $\hat{d}_{mk2} = \hat{A}\hat{d}'_{mk}$. For the items used in reporting scores, test characteristic curves

$$\hat{T}_1(\theta) = \sum_{m=M+1}^{M_1} \sum_{k=1}^{r_m-1} k\hat{P}_m(k|\theta)$$

and

$$\hat{T}_2(\theta) = \sum_{m=M_1+1}^{M_2} \sum_{k=1}^{r_m-1} k\hat{P}_{m2}(k|\theta)$$

are obtained. Here

$$\log[\hat{P}_{m2}(k|\theta)/\hat{P}_{m2}(k-1|\theta)] = D\hat{a}_{m2}(\theta - \hat{b}_{m2} + \hat{d}_{mk2})$$

for $1 \le k \le r_m - 1$ and $\sum_{k=0}^{r_m-1} \hat{P}_{m2}(k|\theta) = 1$. In Administration 1, total scores for items numbered from $M+1$ to M_1 can range from 0 to

$$S_1 = \sum_{m=M+1}^{M_1} (r_m - 1).$$

In Administration 2, total scores for items numbered from $M_1 + 1$ to M_2 can range from 0 to

$$S_2 = \sum_{m=M_1+1}^{M_2} (r_m - 1).$$

Consider a total score s for Administration 2. In true-score equating, an s_2 of 0 in Administration 2 is linked to $s_1 = 0$ in Administration 1, while $s_2 = S_2$ in Administration 2 is linked to $s_1 = S_1$ in Administration 1. If $0 < s_2 < S_2$, then s_2 in Administration 2 is linked to $\hat{T}_1(\theta_2)$, where $\hat{T}_2(\theta_2) = s_2$.

If the model assumptions employed all hold, it is readily shown that all estimates developed in this section have suitable properties for application of the jackknife. Thus the example is appropriate for the jackknifing methods developed in section 1.7. As previously noted, there is some complication to the extent that items are not, in practice, selected at random in typical educational tests. Thus some care will be needed in discussing the effect of selection of anchor items.

It should be noted that the standard errors determined by jackknifing apply whether the model assumptions used in linking are true or not. Parameters still have asymptotic means, but their interpretation is more complex (Haberman, 2007).

3 Example

To illustrate methodology, data from two sections of an assessment are considered for two administrations. In the first section, to be termed section 1, 42 items are used to score examinees, and 28 anchor items are employed, so that $M = 28$, $M_1 = 70$, and $M_2 = 112$. In the second section, to be termed section 2, 34 items are used to score examinees, and 17 anchor items are used, so that $M = 17$, $M_1 = 51$, and $M_2 = 85$. Total sample sizes are about 6,000 for Administration 1 and 8,000 for Administration 2. About 1,600 examinees in each administration receive the anchor items for section 2. In the case of section 1, about 3,100 examinees receive the anchor items for Administration 2, and about 1,600 receive the anchor items for Administration 1. As previously noted, for jackknifing of examinees, 120 disjoint subsets are employed.

Table 1 provides estimates and standard errors for parameters A and B for the two sections. In these computations, the common items are assumed to be given. Results for a model in which the common items are randomly drawn differ for the two sections. The effects of item selection are examined by removal of one anchor item at a time. In section 1, the ratio $F_{GS} = \hat{\sigma}_{\gamma}^2/\hat{\Delta}_{GS}$ of (63) is 2.44 for estimates of A and 4.01 for estimates of B. Given that M is 28 and k is 120, both F statistics are very highly significant, so that σ_{ν}^2 and σ_{γ}^2 appear to be positive, and the effect of anchor selection is a concern. With anchor sets regarded as random, the estimated asymptotic standard deviation of A is increased to 0.0323, and the estimated asymptotic standard deviation of B is increased to 0.0330. These estimates are considerably larger than the customary estimated asymptotic standard errors, so there is cause for concern about the effects of item selection. In section 2, the respective F_{GS} statistics are 0.81 and 1.18, and M is now 17, so that no clear

evidence is present that selection of anchor items has an effect.

Table 1

Estimated A and B Parameters and Estimated Asymptotic Standard Errors

Parameter	Section	Estimate	Standard error
A	1	0.989	0.024
B	1	-0.093	0.022
A	2	0.956	0.029
B	2	-0.086	0.029

Table 2 provides results for conversions of total item scores from Administration 2 to total item scores from Administration 1 for section 1. Table 3 provides the corresponding result for section 2. For section 1, there is an appreciable effect of anchor selection, but the basic result remains that standard errors can be as large as about a third of a raw score point. In section 2, selection of anchor items does not have an obvious effect, so only the conventional results are provided for the grouped jackknife. The standard errors are roughly comparable to those for section 1. Impact of the standard errors in practice depends on the choice of raw-to-scale conversion used in the testing program, on the standard deviation of scaled test scores, and on whether the assessment is applied to individual examinees or to groups of examinees. Reasons for the variability of results due to the specific anchor items selected in section 1 require further investigation. The general issue raised is that it is possible in practice for the selection of anchor items to have an appreciable effect on the variability of equating results.

4 Conclusions

The analysis in this report indicates that jackknifing may be employed both to examine sampling variability in equating and to analyze sensitivity of equating results to anchor selection. The approach used is very widely applicable to equating studies. It would also apply were alternative linking procedures applied such as concurrent calibration or mean and sigma methods (Hambleton et al., 1991). The approach may also be used when the general partial credit model is replaced by the partial credit model with a_m constant for all m (Muraki, 1997). It is also quite possible to apply the approach to observed-score equating methods such as kernel equating (von

Table 2

Estimated Conversions of Total Item Score for Section 1

		Stand. err. for anchor				Stand. err. for anchor	
Score	Estimate	Fixed	Random	Score	Estimate	Fixed	Random
0	0.000	0.000	0.000	23	20.868	0.221	0.279
1	1.255	0.113	0.135	24	21.792	0.213	0.277
2	2.338	0.165	0.200	25	22.729	0.205	0.276
3	3.343	0.202	0.246	26	23.680	0.197	0.277
4	4.300	0.230	0.280	27	24.645	0.192	0.280
5	5.226	0.251	0.304	28	25.625	0.188	0.285
6	6.130	0.269	0.323	29	26.621	0.186	0.292
7	7.015	0.283	0.338	30	27.633	0.185	0.300
8	7.888	0.292	0.348	31	28.663	0.186	0.308
9	8.752	0.301	0.355	32	29.712	0.190	0.317
10	9.608	0.304	0.357	33	30.781	0.193	0.326
11	10.458	0.305	0.358	34	31.871	0.198	0.333
12	11.306	0.305	0.356	35	32.985	0.202	0.339
13	12.152	0.302	0.352	36	34.123	0.205	0.343
14	12.999	0.301	0.347	37	35.287	0.206	0.342
15	13.847	0.293	0.340	38	36.476	0.206	0.338
16	14.699	0.288	0.333	39	37.690	0.202	0.326
17	15.555	0.281	0.325	40	38.924	0.194	0.307
18	16.418	0.271	0.315	41	40.168	0.180	0.279
19	17.289	0.262	0.307	42	41.406	0.157	0.238
20	18.168	0.252	0.298	43	42.612	0.126	0.185
21	19.057	0.243	0.292	44	43.760	0.085	0.117
22	19.957	0.232	0.285	45	45.000	0.000	0.000

Davier, Holland, & Thayer, 2004). The example, as is common in textbook discussions, considers a simple linking of one form to another; however, the methodology is also suitable for examination

Table 3

Estimated Conversions of Total Item Score for Section 2

Score	Estimate	Stand. err. for anchor fixed	Score	Estimate	Stand. err. for anchor fixed
0	0.000	0.000	18	15.360	0.228
1	0.821	0.118	19	16.250	0.215
2	1.617	0.178	20	17.153	0.204
3	2.418	0.222	21	18.068	0.195
4	3.240	0.259	22	18.997	0.189
5	4.082	0.288	23	19.942	0.184
6	4.940	0.310	24	20.904	0.183
7	5.807	0.324	25	21.886	0.183
8	6.678	0.331	26	22.894	0.187
9	7.551	0.333	27	23.935	0.193
10	8.421	0.330	28	25.016	0.199
11	9.288	0.323	29	26.148	0.205
12	10.152	0.313	30	27.341	0.209
13	11.014	0.300	31	28.602	0.208
14	11.876	0.287	32	29.939	0.199
15	12.739	0.271	33	31.348	0.175
16	13.606	0.257	34	32.803	0.122
17	14.479	0.242	35	34.000	0.000

of a much more complex sequence of test forms that are linked through many different sets of anchor items.

Consideration of both the sampling variability and the variability of equating results with respect to anchor selection is important in any assessment of the effectiveness of equating for a testing program. It is clearly important for the variability of equating results to be small relative to the measurement error for individual examinees. For example, consider the following case. With the conversion associated with an arbitrarily large sample of examinees and an arbitrarily

large selection of different anchor sets, the standard deviation of an examinee's equated score on a form has a standard deviation of 5, and 0.84 is the form reliability of this score. Thus the standard error of measurement is 2. Suppose that use of finite samples to compute equating functions and use of one of many possible anchor sets results in a random scoring error for the examinee with mean 0 and standard deviation 1, and suppose that the random scoring error is uncorrelated with the examinee's error of measurement based on the ideal conversion. Then the effective standard error of measurement is $(2^2 + 1^2)^{1/2} = 2.236$ rather than 2. The effective standard deviation is $(5^2 + 1^2)^{1/2} = 5.099$, and the effective reliability is reduced to $1 - (2^2 + 1^2)/(5^2 + 1^2) = 0.808$. The equating error impact can be far more important when a group of examinees is studied. Consider a sample of 100 randomly selected examinees for the form under study. For these examinees, the standard deviation of the mean equated scores for the ideal conversion is 0.5, and the standard deviation of the mean error of measurement is 0.2, so the reliability of the estimated mean score remains 0.84. On the other hand, it is quite possible that the mean random scoring error has essentially the same distribution as the random scoring error for a single examinee, so that 1 remains the standard deviation of the mean random scoring error. Thus the effective reliability of the mean equated score is now only $1 - (0.2^2 + 1^2)/(0.5^2 + 1^2) = 0.168$.

The sensitivity of equating to the selection of anchor items is particularly important, for this problem does not become unimportant even when sample sizes are very large. As a consequence, it is of great importance that equating procedures be investigated for robustness to selection of anchor items. The approach in this report provides an appropriate method of investigation for both sampling errors and errors due to selection of anchor items. A general appreciation of the stability of equating results with respect to sample size and anchor selection requires a more comprehensive investigation of equating results from a substantial number of test administrations for a variety of testing programs. Such data can indicate the magnitude of variability commonly encountered and can suggest circumstances which lead to higher or lower variability.

References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions.* New York: Dover.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*, 443–459.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*, 290–302.

Cohen, J., Johnson, E., & Angeles, J. (2001). *Estimates of the precision of estimates from NAEP using a two-dimensional jackknife procedure.* Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1-26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Haberman, S. J. (2007). *The information a test provides on an ability parameter* (ETS Research Rep. No. RR-07-18). Princeton, NJ: ETS.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Miller, R. G. (1964). A trustworthy jackknife. *The Annals of Mathematical Statistics*, *35*, 1594-1605.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.

Scheffé, H. (1959). *The analysis of variance.* New York: John Wiley.

Shao, J. (2003). *Mathematical statistics* (2nd ed.). New York: Springer.

Shao, J., & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, *17*, 1176–1197.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples [Abstract]. *The Annals of Mathematical Statistics*, *29*, 561.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer.